https://doi.org/10.1038/s41587-023-02100-3

Scalable, accessible and reproducible reference genome assembly and evaluation in Galaxy

The Earth BioGenome Project aims to produce reference genomes for all ~1.8 million known eukaryotic species over the next decade[1–4]. Achieving this goal will require the current pace of reference genome production to increase by at least two orders of magnitude[1]. Automation of the assembly process with a pipeline that is widely accessible to any research group will be required to achieve this speed-up. Enabling this goal requires sustained effort in three major areas: genome assembly optimization and best-practice development, computational infrastructure provisioning, and dissemination and training.

To optimize the assembly process and devise best practices, we combined the expertise of two projects—the Vertebrate Genomes Project (VGP) and the European Reference Genome Atlas (ERGA). The VGP is a collaborative effort to generate reference genomes for all ~70,000 vertebrate species[5]. In the past 5 years, the VGP has released hundreds of new reference genomes supported by the development of automated assembly tools and workflows[1,5]. The ERGA is a pan-European scientific initiative to generate reference genomes for all ~200,000 European eukaryote species, many of which are on the International Union for Conservation of Nature Red List of species at risk of extinction[2].

Advancing from the prior VGP work, originally on the DNAnexus platform (Supplementary Note, section 1.1), we developed a pipeline within the Galaxy ecosystem[6] that combines Pacific Biosciences (PacBio) high-fidelity (HiFi) reads with long-distance information from Hi-C maps and/or optical maps to generate nearly complete assemblies (Supplementary Note 1.3). The pipeline further uses Hi-C or whole-genome sequence data from parents to produce chromosomal-level or whole-genome-level phased genomes, respectively. To streamline the assembly process and ensure quality, the pipeline includes extensive quality control (QC) functions at every step (Supplementary Fig. 1 and Supplementary Note, section 2.1). We suggest at least 30× PacBio HiFi coverage, and up to 60× coverage to accurately assemble highly repetitive regions, as well as 30× Hi-C coverage per haplotype. This is important to ensure a uniform read distribution during the random Poisson sampling process of whole-genome sequencing[7].

Galaxy allows users to execute complex workflows on thousands of datasets and terabytes of data either via a graphical user interface or programmatically via application programming interface (API) scripts[8]. Major global Galaxy instances in the United States (https://usegalaxy.org), the European Union (http://usegalaxy.eu) and Australia (https://usegalaxy.org.au) are freely accessible to researchers worldwide and supported by public cloud infrastructures so that users are not required to install any tools or procure any infrastructure. Galaxy can also be installed locally to use existing high-performance

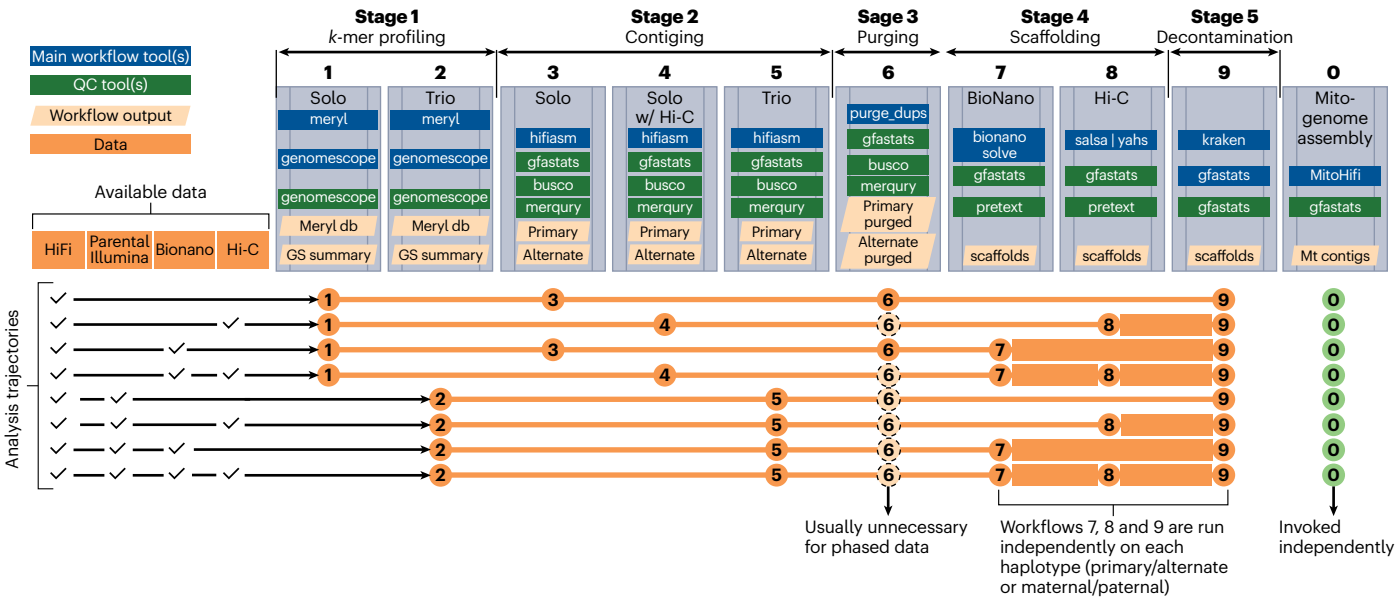

Fig. 1 | VGP–Galaxy assembly pipeline (version 2.1) consists of 10 workflows that can be combined into 8 analysis trajectories depending on the combination of input data. A decision on whether to invoke workflow 6 is based on the analysis of QC output of workflows 3, 4 or 5 (see Supplementary Information for full explanation). Thicker lines connecting workflows 7, 8 and 9 reflect the fact that these workflows are invoked separately for each phased assembly (once for maternal and once for paternal).

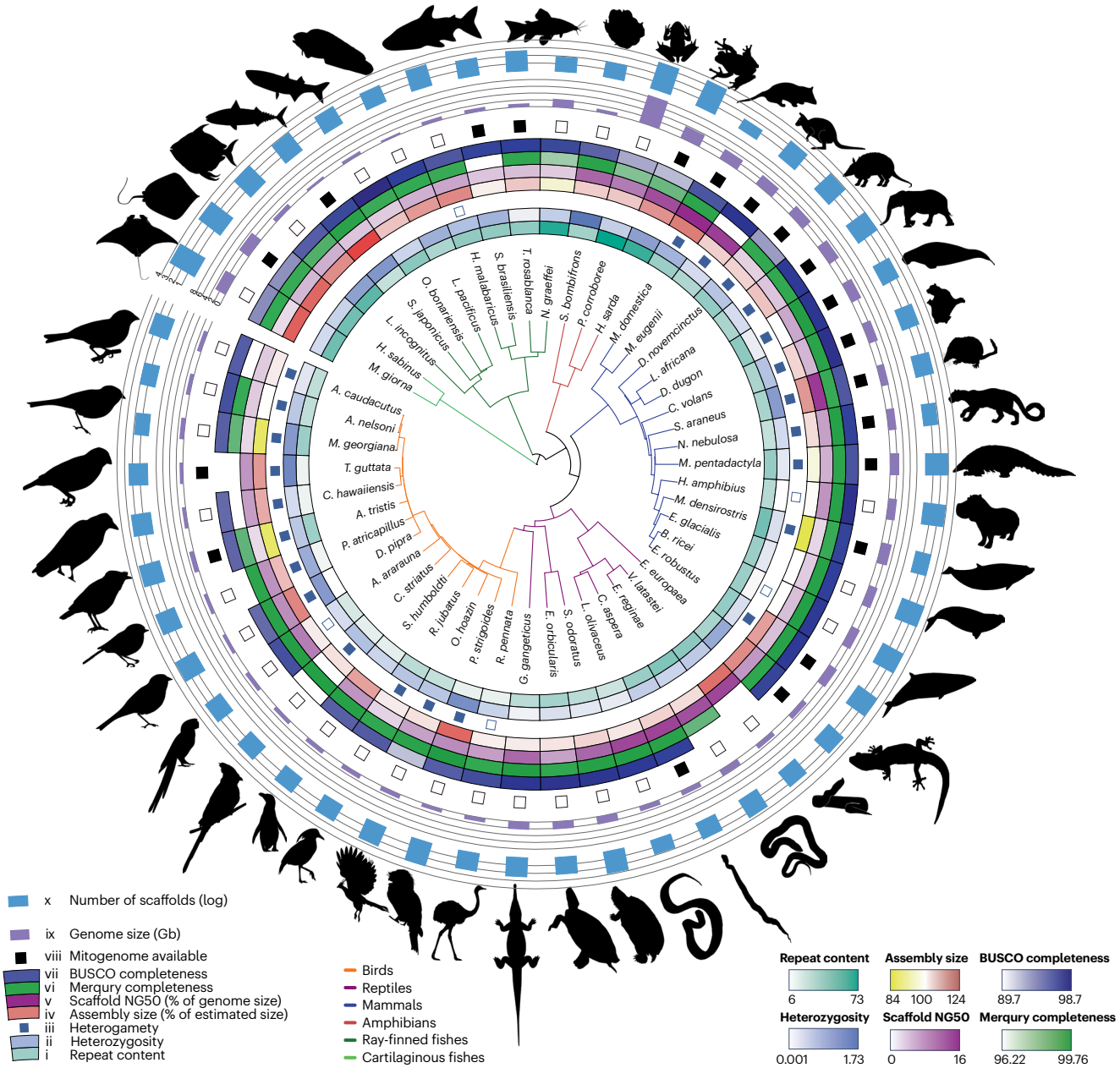

Fig. 2 | Phylogenetic tree and assembly statistics of genomes assembled using the VGP–Galaxy assembly pipeline. From the innermost circle to the outermost circle: (i) repeat content; (ii) heterozygosity; (iii) heterogamy: individuals with two identical sex chromosomes (white) or two different sex chromosomes (blue); (iv) assembly size in percentage of the genome size estimated by Genomescope; (v) scaffold NG50 in % of estimated genome size; (vi) Merqury completeness of both haplotypes; (vii) BUSCO completeness: presence of orthologous genes present and complete compared to the set expected in vertebrates; (viii) mitogenome assembled and available (black); (ix) genome size in gigabytes, with lines at 9, 2, 3, 4, 6 and 8 Gb; (x) number of scaffolds in log scale, with lines at 1 (10 scaffolds), 2 (100 scaffolds), 3 (1,000 scaffolds) and 4 (10,000 scaffolds).

computing (HPC) systems and configured to access heterogeneous, geographically distributed storage and computing resources[9].

The resulting VGP–Galaxy assembly pipeline is organized into 10 Galaxy workflows (Fig. 1; Supplementary Note, section 2.1) to account for different combinations of input data and stages of the assembly process. We systematically evaluated several scaffolding approaches, resulting in best-practice workflows using Hi-C and/or Bionano optical mapping data. We further implemented a dedicated mitogenome assembly pipeline to validate species identification and provide mitochondrial reference assemblies[10,11].

We also developed a decontamination workflow to remove exogenous sequences (e.g., viral and bacterial sequences), as well as mitochondrial artifacts that are often present in draft assemblies, as required for submission to public archives (Supplementary Note, section 2.2.4).

We first tested the automated workflows on the assembly of a reference genome of zebra finch (*Taeniopygia guttata*), for which a wide variety of genomic sequencing data types are available. This led to the development of three types of assembly trajectories (Fig. 1 and Supplementary Table 1): solo assembly (workflows 1, 3, 6 and 9; Fig. 1) using PacBio HiFi data for single individuals; Hi-C assembly (workflows 1, 4, 8 and 9) obtained by adding Hi-C data for phasing and scaffolding the contigs; and trio assembly (workflows 2, 5, 8 and 9) produced by using Illumina short-read data from parents for haplotype phasing (Fig. 1 and Supplementary Table 1).

To validate the pipeline, we used 51 vertebrate datasets for which PacBio HiFi and Hi-C data were available. We compared these assemblies against 19 previous PacBio continuous long read–based genomes of similar size and complexity to confirm and extend the improvements to HiFi technology over continuous long-read methods reported previously[12] (Fig. 2, Supplementary Table 5, Supplementary Fig. 6).

Given the improved haplotype resolution that resulted from adding Hi-C data, even for large (~4.3 Gbp), repeat-rich genomes, we recommend Hi-C Hifiasm phasing when parental data are not available. It is now possible to use well-tested kits as long as samples have been preserved properly (fresh frozen and without DNA and RNA preservatives that protect DNA but reduce protein crosslinks). For use with difficult-to-obtain samples, we have included pipeline options that do not require Hi-C data (Fig. 1).

Although all genome assemblies reported here are for vertebrates, the above principles and our pipeline can be applied to other animal, plant or fungal genomes by modifying a few parameters such as, for example, BUSCO clades necessary for accurate QC reporting (Supplementary Methods, section 3.3).

Our approach is designed to be useful across the full spectrum of user skill levels and analysis scenarios. For this purpose, we created dedicated tutorials distributed via the Galaxy Training Network portal[13] that include extended versions and that collectively provide an in-depth overview of the assembly process, as well as a streamlined tutorial designed to facilitate immediate use of the workflows[14].

Our future work will focus on the continuous maintenance of the pipeline to improve its efficiency and scalability, automation of the curation process, incorporation of ultra-long-read data and development of effective genome annotation procedures.

To increase the robustness of the pipeline, we are developing additional workflows to take advantage of Oxford Nanopore Technologies (ONT) data, and particularly of ultra-long (UL) reads (>100 kb). These workflows use HiFi/UL hybrid assembly tools such as Verkko[15] and the HiFi+UL version of Hifiasm[16], both of which we integrated into Galaxy. Each technology complements missing information from the other, with ONT reads being less accurate and HiFi reads being shorter and underperforming on certain genomic patterns, leading to sequencing bias that could affect specific taxa (Supplementary Fig. 14). This integration of complementary sequencing technologies will make our pipeline even more effective at generating complete and accurate reference genomes.

Data availability

The workflows, their description and instructions on how to use them can be found at https://galaxyproject.org/projects/vgp/workflows/. The requisite tools are installed on usegalaxy.org and usegalaxy.eu, and are in the process of being installed on usegalaxy.org.au. These genomes were supported by collaborators of the VGP and ERGA, and the QC analyses reported here to test the VGP Galaxy pipeline do not release those that are under specific embargo policies for genome-wide analyses (e.g., https://genome10k.ucsc.edu/data-use-policies/). New genome assemblies are available in the GenomeArk repository: https://www.genomeark.org/. After manual curation, the assemblies are submitted to the US National Center for Biotechnology Information (NCBI) under the BioProject Vertebrate Genome Project: https://www.ncbi.nlm.nih.gov/bioproject/489243 [17].

Delphine Larivière[1,28], Linelle Abueg[2,28], Nadolina Brajuka[2,28], Cristóbal Gallardo-Alba[3], Bjorn Grüning[3], Byung June Ko[4], Alex Ostrovsky[5], Marc Palmada-Flores[6], Brandon D. Pickett[7], Keon Rabbani[8], Agostinho Antunes[9,10], Jennifer R. Balacco[2], Mark J. P. Chaisson[8], Haoyu Cheng[11,12], Joanna Collins[13], Melanie Couture[2], Alexandra Denisova[14], Olivier Fedrigo[2], Guido Roberto Gallo[15], Alice Maria Giani[16], Grenville MacDonald Gooder[2], Kathleen Horan[2], Nivesh Jain[2], Cassidy Johnson[2], Heebal Kim[4,17,18], Chul Lee[18,19], Tomas Marques-Bonet[6,20,21,22], Brian O'Toole[2], Arang Rhie[7], Simona Secomandi[23], Marcella Sozzoni[24], Tatiana Tilley[2], Marcela Uliano-Silva[25], Marius van den Beek[1], Robert W. Williams[26], Robert M. Waterhouse[27], Adam M. Phillippy[7], Erich D. Jarvis[2] ✉, Michael C. Schatz[5] ✉, Anton Nekrutenko[1] ✉ & Giulio Formenti[2] ✉

[1]Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA, USA. [2]Vertebrate Genome Laboratory, The Rockefeller University, New York, NY, USA. [3]Bioinformatics Group, Department of Computer Science, Albert-Ludwigs University Freiburg, Freiburg, Germany. [4]Department of Agricultural Biotechnology and Research Institute of Agriculture and Life Sciences, Seoul National University, Seoul, Republic of Korea. [5]Departments of Biology and Computer Science, Johns Hopkins University, Baltimore, MD, USA. [6]Department of Medicine and Life Sciences (MELIS), Institut de Biologia Evolutiva, Universitat Pompeu Fabra-CSIC, Barcelona, Spain. [7]Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [8]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA, USA. [9]CIIMAR/CIMAR, Interdisciplinary Centre of Marine and Environmental Research, University of Porto, Porto, Portugal. [10]Department of Biology, Faculty of Sciences, University of Porto, Porto, Portugal. [11]Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA, USA. [12]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [13]Wellcome Sanger Institute, Cambridge, UK. [14]Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia. [15]Department of Biosciences, University of Milan, Milan, Italy. [16]BMRI, Weill Cornell Medical College, New York, NY, USA. [17]eGnome, Inc., Seoul, Republic of Korea. [18]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea. [19]Laboratory of Neurogenetics of Language, The Rockefeller University, New York, NY, USA. [20]Catalan Institution of Research and Advanced Studies

(ICREA), Barcelona, Spain. [21]CNAG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. [22]Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Spain. [23]Department of Biological Sciences, University of Cyprus, Nicosia, Cyprus. [24]University of Florence, Department of Biology, Florence, Italy. [25]Tree of Life, Wellcome Sanger Institute, Cambridge, UK. [26]Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN, USA. [27]Department of Ecology & Evolution and Swiss Institute of Bioinformatics, University of Lausanne, Lausanne, Switzerland. [28]These authors contributed equally: Delphine Larivière, Linelle Abueg, Nadolina Brajuka
✉e-mail: ejarvis@rockefeller.edu; mschatz@cs.jhu.edu; anton@nekrut.org; gformenti@rockefeller.edu

Published online: 26 January 2024

References

1. Hotaling, S., Kelley, J. L. & Frandsen, P. B. *Proc. Natl Acad. Sci. USA* **118**, e2109019118 (2021).
2. Formenti, G. et al. *Trends Ecol. Evol.* **37**, 197–202 (2022).
3. Theissinger, K. et al. *Trends Genet.* **39**, 545–559 (2003).
4. Lewin, H. A. et al. *Proc. Natl Acad. Sci. USA* **119**, e2115635118 (2022).
5. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. *Genome Biol.* **21**, 245 (2020).
6. Galaxy Community. *Nucleic Acids Res.* **50**, W345–W351 (2022).
7. Lander, E. S. & Waterman, M. S. *Genomics* **2**, 231–239 (1988).
8. Bray, S. & Maier, W. Automating Galaxy workflows using the command line. *Galaxy Training Network* (2023).
9. Galaxy Community. Galaxy Server administration. *Galaxy Training Network* https://github.com/galaxyproject/training-material (2019).
10. Formenti, G. et al. *Genome Biol.* **22**, 120 (2021).
11. Uliano-Silva, M. et al. *BMC Bioinform.* **24**, 288 (2023).
12. Wenger, A. M. et al. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
13. Batut, B. et al. *Cell Syst.* **6**, 752–758.e1 (2018).
14. Lariviere, D., Ostrovsky, A., Gallardo, C., Pickett, B. & Abueg, L. VGP assembly pipeline - short version. *Galaxy Training Network* (2023); https://gxy.io/GTN:T00040
15. Rautiainen, M. et al. *Nat. Biotechnol.* **41**, 1474–1482 (2023).
16. Cheng, H., Asri, M., Lucas, J., Koren, S. & Li, H. Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2306.03399 (2023).
17. BioProject Vertebrate Genome Project. *NCBI BioProject* PRJNA489243 (accessed 18 January 2024); https://www.ncbi.nlm.nih.gov/bioproject/489243

Acknowledgements

We thank Yagoub Adam, Tyler Alioto, Jun Aruga, Diego De Panis, Sagane Dind, Diego Fuentes, Shilpa Garg and Jèssica Gómez for contributing to the initial implementation during ELIXIR Biohackathon 2021. We also thank Nate Jue for help testing and developing the pipeline tutorials and Andrea Guarracino for their useful comments to the manuscript. This work was supported in part by the Intramural Research Program of the US National Human Genome Research Institute (NHGRI), the US National Institutes of Health (NIH) and the Howard Hughes Medical Institute (HHMI). The authors are grateful to the broader Galaxy community for their support and software development efforts. This work is funded by NIH grants U41 HG006620, U24 HG010263, U24 CA231877 and U01CA253481, along with US National Science Foundation grants 1661497, 1758800 and 2216612. The work was also supported in part by The Human Frontier Science Program (HFSP) RGP0025/2021, the Swiss National Science Foundation (SNSF) grants 202669 and 198691, the Swiss State Secretariat for Education, Research and Innovation (SERI) grant 22.00173 and Horizon Europe under the Biodiversity, Circular Economy and Environment program (REA.B.3, BGE 101059492). Usegalaxy.eu is supported by German Federal Ministry of Education and Research grants 031L0101C and de.NBI-epi to B.G. Computational resources are provided by the Advanced Cyberinfrastructure Coordination Ecosystem (ACCESS-CI), Texas Advanced Computing Center, and the JetStream2 scientific cloud.

Author contributions

D.L. built the assembly pipeline with support from G.F., L.A., C.G.-A., B.G., A.O., H.C., M.U.-S., B.D.P., A.R., M.v.d.B. and the VGP assembly working group. L.A., A.D., G.R.G., A.M.G., G.M.G., N.J., C.J., B.O., S.S., M.S. and T.T. generated one or several assemblies used in the analyses. B.J.K., K.R. and M.J.P.C. validated the zebra finch assemblies. J.C. performed the manual curation on the zebra finch assembly. L.A. assembled and evaluated the mitochondrial genomes. N.B. established the decontamination pipeline and performed the contamination analyses. N.B. and M.P.-F. compared the scaffolding strategies. A.N. performed the analyses on XBP1. C.G.-A. and B.D.P. developed the training material with support from the user community. K.H. and M.C. sourced and arranged for sample procurement for species in this study. J.R.B., N.J., T.T., B.O'T., O.F., C.L., H.K., T.M.-B. and R.M.W. generated the PacBio and Hi-C data. G.F., M.C.S., A.N., A.M.P. and E.D.J. conceived the study and drafted the manuscript. All authors, including A.A. and R.W.W., contributed to writing and editing the manuscript and approved it.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-023-02100-3.